Metabarcoding and Metagenomics 6: 27–39
DOI 10.3897/mbmg.6.77704

Research Article

Comparing PCR-generated artifacts of different polymerases for improved accuracy of DNA metabarcoding

Satoshi Nagai[1], **Sirje Sildever**[1,2], **Noriko Nishi**[3], **Satoshi Tazawa**[3], **Leila Basti**[4], **Takanori Kobayashi**[1], **Yoshizumi Ishino**[5]

1 *Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, 2-12-4 Fukuura, Kanazawa-ku, Yokohama, Kanagawa 236-8648, Japan*

2 *Department of Marine Systems, Tallinn University of Technology, Akadeemia tee 15A, 12618 Tallinn, Estonia*

3 *AXIOHELIX Co. Ltd, 1-12-17 Kandaizumicho, Chiyoda-ku, Tokyo 101-0024, Japan*

4 *Department of Ocean Sciences, Faculty of Marine Resources and Environment, Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato-ku, Tokyo 108-8477, Japan*

5 *Department of Bioscience and Biotechnology, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan*

Corresponding author: Satoshi Nagai (snagai@affrc.go.jp)

Academic editor: Alexander Weigand | **Received** 7 November 2021 | **Accepted** 25 January 2022 | **Published** 21 February 2022

Abstract

Accuracy of PCR amplification is vital for obtaining reliable amplicon-sequencing results by metabarcoding. Here, we performed a comparative analysis of error profiles in the PCR products by 14 different PCR kits using a mock eukaryotic community DNA sample mimicking metabarcoding analysis. To prepare a mock eukaryotic community from the marine environment, equal amounts of plasmid DNA from 40 microalgal species were mixed and used for amplicon-sequencing by a high-throughput sequencing approach. To compare the differences in PCR kits used for this experiment, we focused on the following seven parameters: 1) Quality, 2) Chimera, 3) Blast top hit accuracy, 4) Deletion, 5) Insertion, 6) Base substitution and 7) Amplification bias amongst species. The results showed statistically significant differences ($p < 0.05$) for all of the seven parameters depending on the PCR kits used. These differences may result from the different DNA polymerases included in each kit, although the result can also be influenced by PCR reaction conditions. Simultaneous analysis of several parameters suggested that kits containing KOD plus Neo (TOYOBO) and HotStart Taq DNA polymerase (BiONEER, CA, US) at the annealing temperature of 65 °C displayed better results in terms of parameters associated with chimeras, top hit similarity and deletions.

Key Words

chimeric sequences, DNA polymerase, high throughput sequencing, metabarcoding, PCR-generated artifacts, 18S rRNA gene

Introduction

DNA polymerases are widely used for DNA manipulation *in vitro*, including DNA cloning, sequencing, labelling, mutagenesis and other purposes. The fundamental ability of DNA polymerases to synthesise a deoxyribonucleotide chain is conserved. However, the more specific properties, including thermostability, fidelity (proofreading activity), processivity (persistence of sequential nucleotide polymerisation) and specificity (proportion of non-specific amplification) vary depending on the polymerase. For example, thermostability can range from < 4 to 1380 min at 95 °C amongst DNA polymerases (van Pelt-Verkuil et al. 2008). Some polymerases also have a 3'-5' proofreading activity that corrects occasionally-occurring nucleotide misincorporations during the extension processes. However, some polymerases, for example, *Taq* DNA polymerase, lack proofreading ability, leaving the errors uncorrected (Taberlet et al. 2018). For this reason, the usage of high-fidelity DNA polymerase has been suggested for reducing erroneous

Copyright *Satoshi Nagai et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

sequences obtained by DNA metabarcoding (Oliver et al. 2015; Sze and Schloss 2019). The fidelity of polymerases with proofreading ability may be up to 300 times higher compared to *Taq* polymerase (van Pelt-Verkuil et al. 2008; Potapov and Ong 2017; ThermoFisher Scientific 2021).

Processivity indicates the number of nucleotides added to the DNA sequence during a single binding event (Wang et al. 2004). Polymerases with higher processivity support amplification of long templates and shorter extension time and lower amount of polymerase are needed for successful amplification (Wang et al. 2004). For example, KOD DNA polymerase exhibits excellent processivity, showing a five-fold higher extension rate (100–130 nucleotides/second) and 10–15-fold higher processivity (> 300 bases) than *Pfu* DNA polymerase (Takagi et al. 1997). Specificity of the DNA polymerases reflects the proportion of non-specific amplification, for example, extension of misprimed targets and primer-dimers that can have a notable impact on the yield and sensitivity of target amplification (Chou et al. 1992). For these reasons, protein engineering techniques to create mutant or artificial DNA polymerases have been successfully applied for developing more powerful DNA polymerases, suitable for specific purposes amongst the different kinds of DNA manipulations (Ishino and Ishino 2014). In addition, manufacturers also provide optimised PCR kits containing polymerase and other reagents for efficient amplification of various templates.

The metabarcoding approach using universal primers in Illumina sequencers has become a popular method for environmental DNA analysis in aquatic ecosystems, owing to the development of high-throughput sequencing (HTS) technologies (Amaral-Zettler et al. 2009; Medinger et al. 2010; Edgcomb et al. 2011; Taberlet et al. 2012; Egge et al. 2013, 2015; Majaneva et al. 2015; Sawaya et al. 2019). The significant advantage of applying HTS-based technology for monitoring biodiversity offers the great potential for more precise species identification, based on genetic information, especially for species that are indistinguishable by traditional morphology-based microscopic observation (Rhodes 1998; John et al. 2005). This technology delivers high-throughput performance and allows for the detection of several hundreds of operational taxonomic units (OTUs), including dominant species and/or hidden flora from aquatic ecosystems (Cheung et al. 2010; Nolte et al. 2010; Monchy et al. 2012; Tanabe et al. 2015; Nagai et al. 2016a, b, 2019; Dzhembekova et al. 2017, 2018; Hirai et al. 2017a, b; Sildever et al. 2019).

The resulting HTS data is influenced by various factors from DNA extraction to bioinformatics data analysis (Oliver et al. 2015; Deiner et al. 2018; Nichols et al. 2018; Jeunen et al. 2019; Santoferrara 2019; van der Loos and Nijland 2020; Zaiko et al. 2021). Amongst those factors, PCR amplification can influence the diversity detected and relative sequence abundances obtained by the HTS data (Haas et al. 2011; Brandarriz-Fontes et al. 2015; Kelly et al. 2019). The influence of polymerase choice on the HTS data has been previously investigated in terms of characterisation of PCR-related errors (Quail et al. 2012; Gohl et al. 2016; Oh et al., n.d.), chimera formation (Lahr and Katz 2009), species occurrence and relative sequence abundances (Haas et al. 2011; Brandarriz-Fontes et al. 2015; Oliver et al. 2015; Nichols et al. 2018; Kawato et al. 2021) and community composition and quality of HTS data (Sze and Schloss 2019). PCR-generated artifacts have also been found to increase as species diversity increases (Qiu et al. 2001); thus, amplicon-sequence data of 16S rRNA or 18S rRNA from environmental DNAs may contain several artifacts. Furthermore, in a cross-laboratory experiment, the choice of the polymerase had a consistently significant effect on the metabarcoding data variability explained by the differential performance of the polymerases used (Zaiko et al. 2021).

Therefore, we examined the effect of the selected PCR kits containing polymerase on the accuracy of amplicon-sequence reads of the 18S rRNA gene using the HTS-based technology. A mock sample of a eukaryote community was prepared by equally pooling plasmid DNAs from 40 microalgal species. The frequency of artifacts was compared amongst 14 PCR kits containing polymerase in the following seven parameters, i.e. frequencies of: 1) Quality; 2) Chimera; 3) Blast top hit accuracy; 4) Deletion; 5) Insertion; 6) Base substitution; and 7) Amplification bias amongst species. The results displayed statistically significant differences in the amplicon sequences from different PCR kits. Various studies can utilise metabarcoding analysis and the results generated through the experiments reported here will contribute to the planning of amplicon-based HTS studies and evaluation of the obtained data in terms of PCR kit (polymerase) associated bias.

Materials and methods

Abbreviations

HTS high-throughput sequencing;
OTUs operational taxonomic units.

DNA sample preparation of a mock community

Clonal strains of 40 microalgal species were isolated from plankton blooms in several localities from Japan (Suppl. material 1: Table S1). All strains were maintained in glass test tubes in 6 ml according to Nagai et al. (2008). The clonal strains were individually cultivated and DNAs were extracted from the harvested cells by the procedure reported previously by Nagai et al. (2008). Universal primer pair (TAReuk454FWD1, F: CCAGCASCYGCGGTAATTCC; TAReuk454REV3, R: ACTTTCGTTCTTGATYRA (Stoeck et al. 2010) was used to amplify the V4 hypervariable regions of the 18S rRNA gene. PCR was performed on the thermal cycler in a reaction mixture (25 µl) containing 1.0 µl of template DNA, 0.2 mM of each dNTP, 1× PCR buffer, 1.5 mM Mg^{2+}, 1.0 U of KOD -Plus- ver.2 (TOYOBO) that has intensive 3' → 5' exonuclease activity and 1.0 µM of each primer. The PCR

cycling conditions were as follows: initial denaturation at 94 °C for 3 min; 30 cycles at 94 °C for 15 s, at 56 °C for 30 s and at 68 °C for 40 s. Results of PCR amplification were checked by agarose gel electrophoresis. Cloning of the amplified DNA fragments followed by sequencing was carried out as described by Nagai et al. (2008). All the 40 sequences obtained in this study are available from GenBank (accession numbers in Suppl. material 1: Table S1). The plasmid DNAs containing a DNA fragment amplified from the target species were purified from each *E. coli* transformant cell, cultivated in 2 ml of the LB medium at 37 °C for 18 hours, using a FastGene Plasmid Mini Kit (NIPPON Genetics, Tokyo, Japan) according to the manufacture's instruction. The purified plasmid DNAs were quantified using a Qubit 2.0 Fluorometer (Life technology, Carlsbad, CA, USA) and the concentrations were 4.2–51.0 ng μl^{-1} (22.2 ± 11.0, Mean ± SD, n = 40). The DNA samples were pooled with an equal amount (150 ng) to the final concentration of 16.0 ng μl^{-1} and stored at -30 °C until use.

Paired-end library preparation and MiSeq sequencing

To carry out metabarcoding analysis using the MiSeq 250PE platform (Illumina, USA), the same universal primer pair targeting 18S rRNA gene V4 hypervariable regions (around 415 bp in length; Stoeck et al. 2010) was used for amplification due to the unavoidable restriction of sequence length. The workflow followed the "*16S metagenomic sequencing library preparation: preparing 16S ribosomal gene amplicons for the Illumina MiSeq system*" distributed by Illumina (part no. 15044223 Rev. B). A two-step PCR approach was employed to construct the paired-end libraries. The first-round PCR amplified the target region using primers 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT + 18S rRNA gene (TAReuk454FWD1) and 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT + 18S rRNA gene (TAReuk454REV3).

In this study, fourteen commercially available DNA kits containing polymerase (Table 1) were compared in terms of the PCR-generated artifacts. For the effective comparison of the PCR performances, the mock community was analysed in triplicate for each PCR kit and condition tested. For all the kits, the first PCR was performed using a thermal cycler (PC-808; ASTEC, Fukuoka, Japan) in a reaction mixture (25 µl) containing 1.0 µl template DNA (15 ng); 0.2 mM of each dNTP; 1× PCR buffer; 1.5 mM Mg^{2+}; 1.0 U KOD-Plus-ver. 2; and 0.3 µM of each primer (Suppl. material 1: Table S2). The PCR cycling conditions were as follows: initial denaturation at 94 °C for 3 min, followed by 25 cycles at 94 °C for 15 s, 56 °C for 30 s and 68 °C for 40 sec. For KOD Plus Neo and HotStart Taq DNA Polymerase, 61 °C and 65 °C were also tested as the annealing temperature. All DNA polymerases used in this study were hot-start types and the pre-denaturing conditions were 94–98 °C for 30 sec–15 min. PCR amplification was evaluated by 1.5% agarose gel electrophoresis. The PCR products were purified using an Agencourt AMPure XP (BECKMAN COULTER, Life Sciences, Brea, California, USA) according to the manufacturer's protocol.

The second PCR step was performed with a primer set of 5'-AATGATACGGCGACCACCGAGATCTACAC- 8 bp index -ACACTCTTTCCCTACACGACGC (forward) and 5'-CAAGCA GAAGACGGCATACGAGAT- 8 bp index -GTGACTGGAGTTCAGACGTGTG (reverse). The eight base segments represent dual-index sequences used to recognise each sample; the 5' end-sequences are adapters that allow the final product to bind or hybridise to short oligonucleotides on the surface of the Illumina flow cell and the 3' end-sequences are priming sites for the MiSeq sequencing. The purified first-round PCR product was diluted 5 times with TE buffer and used as a template for the second-round PCR. The second-round PCR was carried out with the same condition as the first-round PCR, except the reaction volume of 50 µl containing 2.0 µl of the diluted PCR product. Eight cycles of the PCR with the annealing temperature of 59 °C were set for all PCR reactions tested in this study (Suppl. material 1: Table S2). PCR amplification was verified by agarose gel electrophoresis and the PCR products were purified using an Agencourt AMPure XP (BECKMAN COULTER). The amplified PCR products were quantified and the indexed second PCR products were pooled in equal concentrations and stored at -30 °C until used for sequencing.

When performing high throughput sequencing (HTS), a spike-in of the control library (Illumina standard product: PhiX V3 Control Library) was mixed with the pooled DNA library to improve the data quality of low diversity samples, such as single PCR amplicons. DNA concentrations of the pooled library and the PhiX DNA were adjusted to 4 nM using the buffer EB (10 mM Tris-HCl, pH 8.5) mixed at a ratio of 7:3.5 µl. The 4 nM library was denatured with 5 µl of fresh 0.1 N NaOH. Using the HT1 buffer (provided by the Illumina MiSeq v. 2 Reagent kit for 2× 250 bp PE), the denatured library (10 µl; 2 nM) was diluted to a final concentration of 12 pM for sequencing on the MiSeq platform.

HTS data treatment processes and operational taxonomic unit picking

Nucleotide sequences were demultiplexed depending on the 5'-multiplex identifier (MID) tag and primer sequences using the default format in MiSeq. The sequences containing palindrome clips longer than 30 bp and homopolymer longer than 9 bp were trimmed from the sequences at both ends. The 3' tails with an average quality score of less than 30 at the end of the last 25 bp window were also trimmed from each sequence. The 5' and 3' tails with an average quality score of less than 20 at the end of the last window were also trimmed from each sequence. Sequences longer than 250 bp were truncated to 250 bp by trimming the 3' tails. The trimmed sequences shorter than 200 bp were filtered out. The demultiplexing and trimming were

Table 1. Overview of PCR kits compared in this study based on manufacturer`s information, numbering follows Figure 1.

Nr.	PCR kit containing polymerase	Manufacturer	Proofreading ability	Proofreading compared to *Taq*	Hot start
1.	KOD-Plus-Ver.2	TOYOBO	Y	80	Y
2–4.	KOD-Plus-Neo	TOYOBO	Y	80	Y
5.	KOD FX Neo	TOYOBO	Y	80	Y
6.	Go Taq Green Master Mix	Promega	N		Y
7.	Type-it Microsatellite PCR Kit	QIAGEN	N		Y
8.	Q5 High-fidelity DNA Polymerase	NEW ENGLAND BioLabs	Y	280	Y
9.	One Taq Hot Start DNA Polymerase	NEW ENGLAND BioLabs	Y[a]	5[b]	Y
10.	KAPA2GRobust HotStart Redy Mix with dye (2X)	KAPA Biosystems	N		Y
11.	KAPA HiFi HotStart Ready Mix	KAPA Biosystems	Y	100	Y
12.	Premix Ex Taq Hot Start Version	TaKaRa	Y[c]	4.5	Y
13.	Top DNA Polymerase	BiONEER	N		Y
14.	Pfu DNA Polymerase	BiONEER	Y	30[d]	Y
15–17.	HotStart Taq DNA Polymerase	BiONEER	N		Y
18.	Platinum SuperFi PCR Master Mix	Invitrogen	Y	300	Y

[a] combination of *Taq* (non-proofreading) and *Vent* (proofreading) polymerases, [b] *Vent* polymerase, [c] combination of *Taq* and exonucelase, [d] Potapov and Ong (2017).

performed using Trimmomatic version 0.35 (http://www.usadellab.org/cms/?page=trimmomatic). The remaining sequences were merged into paired reads using Usearch version 8.0.1517 (http://www.drive5.com/usearch/). In addition, singletons were removed. Subsequently, sequences were aligned using Clustal Omega v. 1.2.0. (http://www.clustal.org/omega/). Multiple sequences were aligned with each other and only sequences that were contained in more than 75% of the read positions were extracted. Filtering and a part of the multiple alignment process were performed using the screen.seqs and filter.seqs commands in Mothur, as described in the Miseq SOP (http://www.mothur.org./wiki/MiSeq_SOP) (Schloss et al. 2011). Erroneous and chimeric sequences were detected and removed using the pre.cluster (diffs = 4) and chimera.uchime (minh = 0.1; http://drive5.com/usearch/manual/uchime_algo.html) (Edgar et al. 2011) commands in Mothur, respectively. Using the unique.seqs command of Mothur, the same sequences were collected into operational taxonomic units (OTUs). The contig sequences were counted as OTUs by count.seqs and used for the subsequent taxonomic identification analysis. Demultiplexed, filtered, but untrimmed sequence data were deposited in the DDBJ Sequence Read Archive under access no. DRA012296.

Taxonomic identification of the OTUs

A subset of the nucleotide database consisting of 40 sequences obtained by sub-cloning of the 18S rRNA gene V4 region from 40 microalgal strains was prepared for a BLAST search. The BLAST search was conducted with NCBI BLAST+ 2.2.30+ (Camacho et al. 2009; Cheung et al. 2010) with default parameters, the subset nucleotide database and all OTU-representative sequences as the query. Subsequently, the taxonomic information was obtained from the BLAST hit with top bitscores for each query sequence and the OTUs of the same top-hit were merged.

Sequence comparisons obtained by PCR with various kits containing different DNA polymerases

The effect of fourteen PCR kits containing polymerase on the accuracy for amplification of 18S-rRNA gene in a mock sample of 40 microalgal species was examined in terms of performance in the following seven parameters: 1) Quality; 2) Chimera; 3) Blast top-hit accuracy; 4) Deletion; 5) Insertion; 6) Base substitution; and 7) Amplification bias amongst species. The calculation of each of the parameters is described in Table 2. Bioinformatics analysis from data processing by Trimmomatic to sequence comparison was performed for each PCR condition and the DNA polymerase (n = 18). In Quality, Trimommatic file outputs were used for the count of reads numbers. Read numbers in each sample were obtained by a simple Linux command to count the line number (LN) (e.g. wc -l *.fastq) and it was divided by four because one sequence information is mentioned with four lines. In Chimera, read numbers before the Chimera check is the same as the read numbers passed in the quality check in Quality. Read numbers after the Chimera check counted by count.seqs was obtained from the Mothur file outputs. Three Perl scripts were made to analyse Blast top hit frequency, Deletion, Insertion, Base substitution, and Amplification bias. Blast XML files were read by blastxml_parser.pl (provided as Suppl. material 2) using Bio::SearchIO module in BioPerl. Results of the BLAST search were processed for each query (OTU) and only the top-hit records were extracted.

The following information: query ID, query length, bit score, top-hit name, top-hit identity, alignment length, query alignment length, query alignment sequence, reference alignment sequence and homology information between the query and reference sequences in the alignment (homology_string) were contained in a text file output. Numbers of top-hit sequences to the reference sequences of 40 species for each query ID were counted in the triplicate samples by count.seqs command in Mothur and merged with the information obtained from Blast XML by merge_tophit_count.pl (provided as Suppl. material 2). The infor-

Table 2. Overview of the seven parameters used for evaluation of PCR kits based on frequencies, PE: paired-end, SD: standard deviation.

Order	Parameter	Definition
A	Quality	Read numbers, passed quality check (the number of reads after merge assemble of PE reads following quality trimming but before chimera check) / raw read numbers in each sample
B	Chimera	Read numbers after Chimera check / read numbers before Chimera check in each sample.
C	Blast top hit accuracy	Numbers of unique hit-sequences to the reference sequences of 40 species / input sequence-numbers in Blast search (= read numbers after Chimera check) in each sample
D	Deletion	Deletion = 1- ((the number of deleted base in OTU1 × the read count of OTU1 + the number of deleted base in OTU2 × the read count of OTU2 + • • • + number of deleted base in OTU40 × read count of OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40))*10^3
E	Insertion	Insertion = 1- ((the number of inserted base in OTU1 × the read count of OTU1 + the number of inserted bases in OTU2 × the read count of OTU2 + • • • + number of inserted base in OTU40 × read count of OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40))*10^3
F	Base substitution	Base substitution = 1- ((the number of base substitution in OTU1 × the read count of OTU1 + the number of base substitution in OTU2 × the read count of OTU2 + • • • + the number of base substitution in OTU40 × read count of OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40))*10^3
G	Amplification bias among species	1-normalized SD of numbers of unique hit-sequences to the reference sequences of 40 species for 40 species

PE, paired-end; SD, standard deviation.

mation was saved as an output in a text file (result_merge.txt). Subsequently, numbers of unique-hit sequences to the reference sequences of 40 species were counted, based on triplicate samples by same_tophit_count_merge.pl (provided as Suppl. material 3). The information was saved as an output in a text file. Removal of sequences containing errors was imperfect after the successive processes of HTS data treatment. Sequences containing different types of errors derived from original ones remained in the following analytical steps. Therefore, these sequences are detected as unique OTUs with the same BLAST top-hit name, but different similarities. BLAST top-hit accuracy and Amplification bias amongst species were calculated using this file. Deletion, Insertion and Base substitution were calculated, based on the following command:

```
awk -F"\t" '!/;/&&NR!=1{ OFS="\t";x[1]=gsub-
("-","-",$8);x[2]=gsub("-","-",$9) ;x[3]= gsub(/\s/,"
",$10);print $1 ,x[1],x[2],x[3]; }' result_merge.txt
```

Namely, to detect Deletion, Insertion and Base substitution, the number of the hyphen '-' in the query alignment sequence, the number of the hyphen '-' in the reference alignment sequence and the number of space between the query and reference sequences in the alignment were counted respectively and the information was saved as an output in a text file. For Amplification bias amongst species, the normalised standard deviation (SD) of numbers of unique hit sequences in 40 species were calculated for each PCR condition in Excel.

To evaluate whether there are statistically significant differences amongst the results from different PCR kits, based on the seven parameters, one-way ANOVA or Kruskal-Wallis rank-sum tests were conducted in R (R Core Team 2021). The assumptions for the tests: normal distribution (ANOVA) and equal variation within groups (ANOVA, Kruskal-Wallis) were also confirmed in R by applying Shapiro-Wilk and Levene's test (package "car", Fox et al. 2020) for normally distributed data or Flinger-Killeen test in the case of non-normal distribution. Tukey Honest Significant Differences method with family-wise confidence level (95%) or Dunn test with the P-value adjusted by the Benjamini-Hochberg method (package "FSA", Ogle et al. 2021) were used as post hoc tests for statistically significant multiple comparisons to obtain more detailed information on the differences amongst the PCR kits tested in this study.

To display performances of PCR kits containing DNA polymerases amongst the investigated parameters intuitively, radar charts were employed. Radar charts represent the integrated performance evaluation of seven parameters in each PCR kit containing polymerase (14 kits) and three different annealing temperature conditions for KOD-Plus-Neo and HotStart Taq DNA Polymerase. The values on the radar charts range from 0–1; if the value is closer to 1, the PCR kit has a better performance in terms of the particular parameter.

Results

All of the seven parameters used for the comparison of the sequence data from different PCR kits displayed statistically significant differences ($p < 0.05$; Table 3). For example, the differences in the results of quality before/after quality trimming amongst different PCR kits and the reaction conditions ranged from 0.29 to 0.41 with higher frequencies indicating higher quality (Fig. 1A). Although the quality amongst different PCR kits varied in a narrow range, a statistically significant difference ($p < 0.05$; Table 3) was detected, related to Premix Ex Taq Hot Start Version (nr. 12; Fig. 1A; Suppl. material 1: Table S3A). For the parameter associated with chimeras, a notable and statistically significant variability amongst the different PCR kits was observed (Fig. 1B; Suppl. material 1: Table S3B). Significantly better performance (lower frequencies of chimeras) was demonstrated by KOD

Table 3. Comparison between different polymerases based on the parameters evaluated.

Parameter	Saphiro-Wilk (W/p value)	Levene's test (F/p value)	Flinger/Killeen (X2/p)	ANOVA (F/p value)	Kruskal-Wallis (X2/p value)
Quality	0.98/0.71	1.64/0.99		3.17/0.002	
Chimera	0.85/**6.645e-06**		9.31/0.93		52.52/**1.702e-05**
Blast top hit accuracy	0.44/ **4.818e-13**		13.23/0.72		52.34/**1.815e-05**
Deletion	0.82/ **1.323e-06**		12.43/0.77		52.181/**1.925e-05**
Insertion	0.91/**0.001**		8.10/0.96		52.10/**1.98e-05**
Base substitution	0.93/**0.003**		12.40/0.77		52.39/**1.782e-05**
Amplification bias	0.57/**2.848e-11**		5.35/0.99		50.68/**3.313e-05**

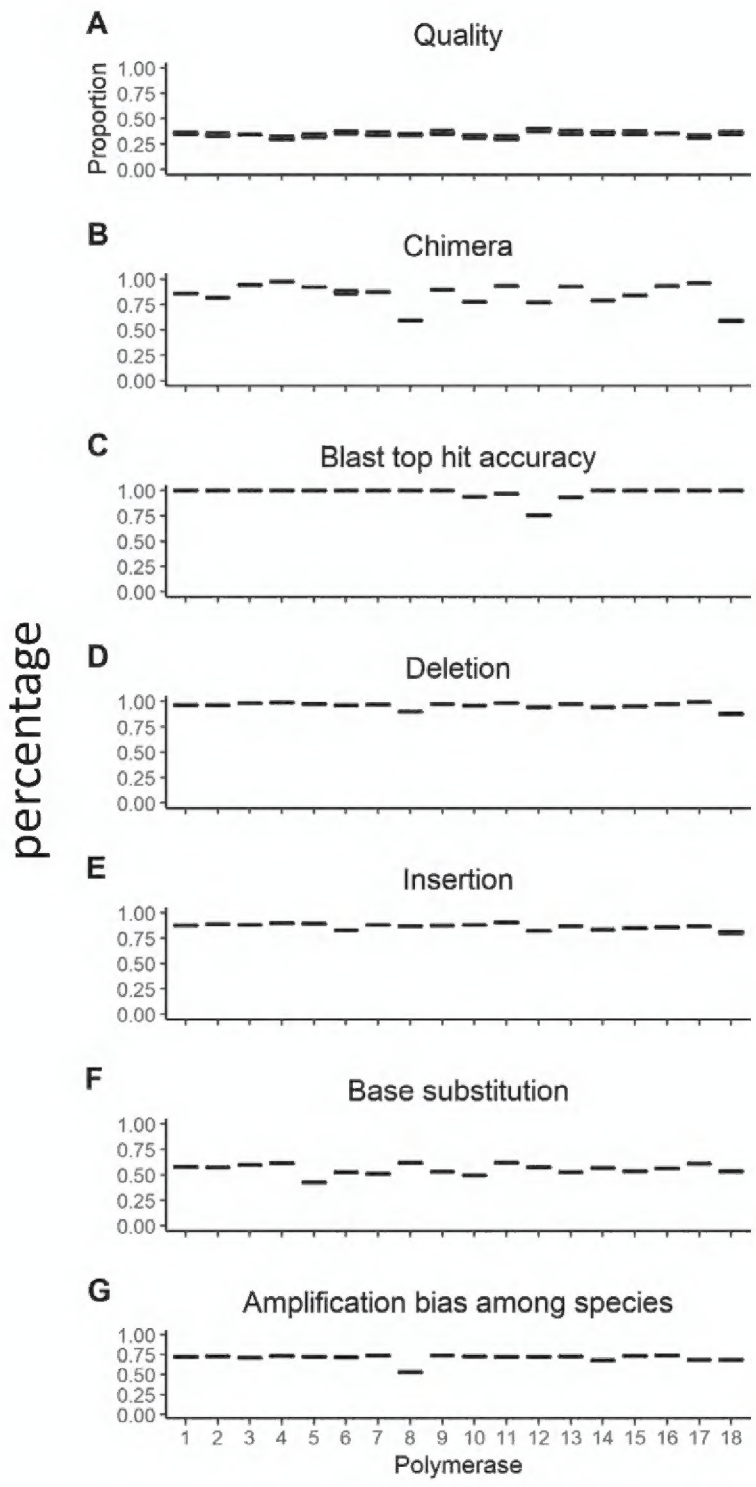

Figure 1. The average performance of different PCR kits containing a polymerase, based on the seven parameters: 1) Quality, 2) Chimera, 3) BLAST top-hit accuracy, 4) Deletion, 5) Insertion, 6) Base substitution and 7) Amplification bias amongst the species. PCR kits: 1: KOD-Plus-Ver.2; 2: KOD-Plus-Neo; 3: KOD-Plus-Neo (61 °C); 4: KOD-Plus-Neo (65 °C); 5: KOD FX Neo; 6: Go Taq Green Master Mix; 7: Type-it Microsatellite PCR Kit; 8: Q5 High-fidelity DNA Polymerase; 9: One Taq Hot Start DNA Polymerase; 10: KAPA2GRobust HotStart Ready Mix with dye (2X); 11: KAPA HiFi HotStart Ready Mix; 12: Premix Ex Taq Hot Start Version; 13: Top DNA Polymerase; 14: Pfu DNA Polymerase; 15: HotStart Taq DNA Polymerase; 16: HotStart Taq DNA Polymerase (61 °C); 17: HotStart Taq DNA Polymerase (65 °C); 18: Platinum SuperFi PCR Master Mix. The numbers shown in the parentheses indicate the annealing temperature in the first-round PCR, otherwise, the first-round PCR was performed at the annealing temperature of 56 °C.

Plus NEO and Hot Start Taq DNA Polymerase with the annealing temperature of 65 °C (nr. 4, 17; Fig. 1B; Suppl. material 1: Table S3B). At the same time, significantly lower performance (higher frequencies of chimeras) was present, when Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (nr. 8, 18; Fig. 1B; Suppl. material 1: Table S3B) were employed.

The performance of reads detected against the unique reference sequences by BLAST search, parameter 3) Blast top-hit accuracy, displayed similarly high performance amongst the majority of the PCR kits, with significantly lower values obtained by KAPA2G Robust HotStart Ready Mix with dye (2X), KAPA HiFi HotStart Ready Mix, Premix Ex Taq Hot Start Version and Top DNA Polymerase (nr. 10- 13; Figs 1C, 2; Suppl. material 1: Table S3C).

In terms of deletions and insertions, the variability amongst the different PCR kits was low (Fig. 1D, E). In contrast, more deletions were detected in the sequences amplified by Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (nr. 8, 18). In the case of insertions, significantly higher bias was detected from the DNA amplified by Go Taq Green Master Mix, Premix Ex Taq Hot Start Version and Platinum SuperFi PCR Master Mix (nr. 6, 12, 18; Fig. 1E, Suppl. material 1: Table S3E). The frequency of base substitutions also varied significantly amongst the PCR kits with the lowest performance by KOD FX Neo and KAPA2G Robust HotStart Ready Mix with dye (2X) (nr. 5, 10; Fig. 1F; Suppl. material 1: Table S3F). Higher performance (the lower numbers of base substitution) was demonstrated by KOD Plus Neo at 65 °C, Q5 High-fidelity DNA Polymerase, KAPA HiFi HotStart Ready Mix and HotStart Taq DNA Polymerase with the annealing temperature of 65 °C (nr. 4, 8, 11, 17; Fig. 1F; Suppl. material 1: Table S3F). The PCR amplification bias amongst the sequences varied significantly in a range from 0.52 to 0.74, with the significantly lowest performance obtained by Q5 High-fidelity DNA Polymerase (nr. 8; Fig. 1G; Suppl. material 1: Table S3G).

Amongst all the parameters evaluated, Premix Ex Taq Hot Start Version displayed significant differences with other PCR kits in four parameters (nr. 12; Fig. 1A–G; Suppl. material 1: Table S3A–G). For other polymerases, the frequency patterns were more variable amongst different parameters (Fig. 1, Suppl. material 1: Table S3). In terms of annealing temperatures, a significant difference for KOD Plus Neo was detected in the frequency of chimeras, top hit similarity, and deletions between 56 °C and 65 °C (nr. 2, 4; Fig. 1B, C; Suppl. material 1: Table

Figure 2. Overview on the performance of the PCR kits/conditions compared amongst seven parameters: 1) Quality, 2) Chimera, 3) Top-hits accuracy (= BLAST top-hit accuracy), 4) Deletion, 5) Insertion, 6) Base substitution and 7) Species bias (= Amplification bias amongst the species). Higher values indicate higher performance in a specific parameter.

S3B, C). For HotStart Taq DNA Polymerase, significant differences between the annealing temperatures were detected for the frequency of deletions at 56 °C and 65 °C (nr. 15, 17; Fig. 1D, Suppl. material 1: Table S3D) as well as for the amplification bias at 61 °C and 65 °C (nr. 16, 17, Fig. 1G; Suppl. material 1: Table S3G).

When considering different parameters together, the majority of PCR kits displayed high performance in two parameters, which were mainly associated with the deletions and top-hit similarities (Fig. 2). Thirteen PCR kits showed the highest results in terms of one parameter that was mainly associated with the top-hit similarity. One PCR kit (Premix Ex Taq Hot Start Version) did not show high performance in any of the parameters, whereas two kits at 65 °C (KOD Plus Neo at 65 °C and Hot Start Taq DNA polymerase at 65 °C) had the best results for the highest number of parameters, mainly associated with chimeras, top-hit similarity and deletions (Fig. 2).

Discussion

We detected significant differences amongst PCR kits for each of the compared parameters when using the mock community sample of marine microeukaryotes. The highest number of significant differences amongst kits was detected in association with BLAST top-hit accuracy, chimeras, insertions and base substitutions, whereas the lowest number of significant differences was detected in terms of the quality of sequences. The low number of significant differences in terms of quality may potentially be explained by similar sequencing depth due to equal concentrations of the pooled samples (Elbrecht and Steinke 2018). Based on all the investigated parameters, the majority of PCR kits performed well in terms of top-hit similarity and deletions.

When analysing the different parameters in more detail, the first parameter displaying a high number of significant differences amongst kits was associated with chimera formation. It varied notably amongst the PCR kits and was the highest in the samples amplified by Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (Fig. 1B). In a previous study, Q5 High-Fidelity Polymerase displayed a notable increase in the percentage of chimeric reads starting from 25 cycles (Gohl et al. 2016) and had the second-highest rate of chimeras in comparison with other kits in 30 cycles (Sze and Schloss 2019). In this study, 25 cycles were used for all comparisons. Thus, the low performance, in terms of chimeras associated with those PCR kits, may result from cycle numbers and could potentially be reduced by reducing the number of cycles.

Previous studies investigating chimera formation and different polymerases report that reducing the amount of DNA template and the number of PCR cycles used (Oh et al., n.d.; Lahr and Katz 2009; Gohl et al. 2016), increasing the extension time (Qiu et al. 2001) and using high-fidelity (Oliver et al. 2015; Sze and Scholss 2019) or high-processivity polymerase (Gohl et al. 2016) can reduce the chimera formation. This is especially important for analysing environmental samples as chimeras appear more frequently in samples with high diversity (Fonseca et al. 2012). In addition, the unequal template concentrations in the environmental samples may contribute to the formation of chimeras (Lahr and Katz 2009). Several bioinformatics tools have been developed to identify and remove chimeras from the HTS data (Edgar et al. 2011; Haas et al. 2011; Quince et al. 2011; Wright et al. 2012). However, as the majority of chimeras may have more than two parent sequences, it might be difficult to detect the chimeras with the software (Lahr and Katz 2009). As an alternative approach, the usage of unique molecular identifiers has been demonstrated to be efficient for detecting chimeras and PCR-related bias in the HTS data (Filges et al. 2019; Fields et al. 2021).

The second parameter displaying a high number of significant differences between the PCR kits was BLAST top-hit similarity. The significant differences were mainly associated with four kits (KAPA2G, KAPA HiFi, Premix Ex Taq and Top DNA polymerase). Interestingly, from those PCR kits, Premix Ex Taq and KAPA2G also displayed lower results in some other parameters, for example, in terms of chimeras and base subtitutions (KAPA2G) and in terms of chimeras, deletions and insertions (Premix Ex Taq). For one of the PCR kits (KAPA2G), the high proportion of errors may be explained by the lack of proofreading ability, whereas the second kit (Premix Ex Taq) has a proofreading ability. Thus, some other parameters or their combination might have a stronger influence on obtaining the correct sequences than the proofreading ability (Brandarriz-Fontes et al. 2015).

The parameters associated with base substitutions also varied notably amongst the PCR kits. As those errors may also arise from sequencing (Glenn 2011; Pfeifer 2017), the results in those parameters may reflect a combination of PCR and sequencing-induced errors. Both the polymerase errors during PCR (substitutions) and sequencing errors (insertion, deletions, substitutions) have been reported as important sources of bias in the HTS data (Patin et al. 2013; Filges et al. 2019). A marginal contribution of polymerase-induced errors to the total errors in the HTS data has been demonstrated, based on the lack of statistically significant differences between the erroneous sequences obtained by proofreading and non-proofreading polymerases (Pfeiffer et al. 2018; Filges et al. 2019). In contrast, a clear pattern in the proportion of erroneous reads in the HTS data between the proofreading and non-proofreading polymerases has also been reported (Nichols et al. 2018). As the results of this study display statistically significant differences amongst different PCR kits and, as all the samples were pooled into one library analysed in a single run, all samples amplified with different PCR kits should experience similar sequencing errors (Lighten et al. 2014). Thus, the sequencing error might be responsible for some variability amongst the PCR kits, but the observed significant differences in terms of insertions, deletion and base substitutions are considered to be more influenced during PCR by the differences between PCR kits and the polymerase included.

Polymerase errors can be influenced by various parameters, such as template DNA, for example, complexity of the gene targeted (single allele or several alleles), presence of repetitive sequences, secondary structure (Cariello et al. 1991; Eckert and Kunkel 1991; Kunkel and Bebenek 2000; Brandarriz-Fontes et al. 2015), the extent of DNA damage from extraction and thermocycling (Eckert and Kunkel 1991; Witzingerode et al. 1997; Potapov and Ong 2017) and by the PCR cycle number (Patin et al. 2013). The influence of the template DNA on the polymerase errors has been exemplified by the proportion of erroneous reads ranging from 8% to 50% between different polymerases when targeting a gene with a single allele (Brandarriz-Fontes et al. 2015). This can partly be explained by the differences in the intrinsic error rate of each polymerase (Hestand et al. 2016). In general, a notably lower proportion of errors associated with insertions and deletions has been reported (< 3%) to

be present in the HTS data compared to base substitution errors (> 96%; Kozich et al. 2013; Potapov and Ong 2017). Thus, the usage of high-fidelity polymerase could help reduce substitution errors originating from PCR as demonstrated in this study by two of the high-fidelity polymerases (Q5, KAPA HiFi). In addition, multi-template PCR, such as the metabarcoding approach using environmental DNA as a template, may be particularly susceptible to PCR biases due to the differences in template DNA sequences and their frequencies, leading to variations in amplification efficiencies (Kalle et al. 2014). Optimisation of PCR mixture component concentration and pH are known to improve the fidelity of polymerases (Ling et al. 1991). However, this should not be a notable influence as the PCR kits tested have been optimised by the manufacturers and, thus, the differences are assumed to reflect the intrinsic differences amongst PCR kits.

Interestingly, the amplification bias amongst species was generally low between the polymerases, except for Q5. The bias results from the high standard deviation between the species counts and should theoretically be similar for all the polymerases as the same mock community was analysed. Stochasticity of the PCR process has been shown to induce skewed sequence representation (Kebschull and Zador 2015). However, in this case, the bias should be greater between the PCR kits. The Q5 polymerase has previously demonstrated a low correlation between the observed and expected oligonucleotide proportions in metabarcoding ($R^2 = 0.44$; Nichols et al. 2018). Thus, the high amplification bias might be influenced by some intrinsic characteristics of this particular polymerase. Furthermore, the Q5 has been reported to prefer sequences with higher GC content (Nichols et al. 2018). However, as the GC content in the mock community varied in a narrow range (40% to 55%, an average of 44%, SD: 3%; data not shown), the differences in GC content are not expected to lead to the observed amplification bias. It has also been found that guanine-rich sequences can cause failure in PCR when proofreading polymerases are used (Zhu et al. 2016). As Q5 has one of the highest proofreading abilities compared to all the polymerases in the kits tested, it could serve as an explanation for the amplification bias amongst the species. However, in the mock community, the guanine content ranged from 24% to 29% (average: 26. 8%, SD: 1%; data not shown) and, thus, it is not certain whether this difference is enough to result in high bias amongst the species. Further experiments are necessary to investigate the potential causes of amplification bias in Q5.

The majority of the kits tested in this study resulted in high values in terms of top-hit accuracy. This coincides with the results of a previous study finding no influence of polymerase choice on the species occurrence data in metabarcoding studies (Nichols et al. 2018). From all the kits analysed and conditions tested, two kits: KOD Plus NEO and HotStart Taq DNA Polymerase, both at 65 °C, displayed the highest values for the highest number of parameters associated with the top-hit similarity, deletions and chimeras. However, it is challenging to suggest the best PCR kit suitable for all metabarcoding studies as the differences in the diversity of the DNA template can also influence the outcome (Qiu et al. 2001). Thus, we recommend considering the importance of different parameters, for example, the seven parameters analyed, while planning a metabarcoding study and testing the suitability of the several PCR kits, for example, the three kits recommended here, for the particular samples.

Conclusions

The purpose of this study was to compare PCR kits containing polymerases to minimise PCR-based amplification artifacts in environmental DNA analysis, based on HTS and metabarcoding. Statistically significant differences amongst PCR kits were detected for all the parameters. The kits displaying significant variability as well as the best results for each parameter varied. The results of the comprehensive analysis visualised by radar charts suggested that KOD plus Neo and HotStart Taq DNA PCR kits with the annealing temperature of 65 °C displayed better performances in the highest number of parameters associated with chimeras, top hit similarity, and deletions. Especially, the higher annealing temperatures reduced chimera formation. However, as the outcome may also be influenced by the template diversity (Qiu et al. 2001) and PCR cycling conditions (Patin et al. 2013), it is challenging to choose a single best PCR kit containing a polymerase that is suitable for all studies. Thus, we recommend using the knowledge generated in this study as a basis for choosing a PCR kit containing a polymerase in combination with further testing and optimisation for particular samples.

Acknowledgements

The authors thank R. Kubota for her assistance with the molecular work. This work was funded by "Technological developments for characterization of harmful plankton in the seawater", Ministry of Agriculture, Forestry and Fisheries, Japan (16808839) [SN]; JST/JICA, Science and Technology Research Partnership for Sustainable Development (JPMJSA1705) [SN]; Japan Society for the Promotion of Science Short-term Postdoctoral Fellowship (PE18028) [SN, SS]; European Regional Development Fund and the programme Mobilitas Plus (MOBTP160) [SS]; Estonian Research Council grant (PSG735) [SS].

References

Amaral-Zettler L, McCliment E, Ducklow H, Huse S (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE 4: e6372. https://doi.org/10.1371/journal.pone.0006372

Brandarriz-Fontes C, Camacho-Sanchez M, Vila C, Vega-Pla L, Rico C, Leonard JA (2015) Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. Scientific Reports 5: e8056. https://doi.org/10.1038/srep08056

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: e421. https://doi.org/10.1186/1471-2105-10-421

Cariello NF, Thilly WG, Swenberg JA, Skopek TR (1991) Deletion mutagenesis during polymerase chain reaction: dependence on DNA polymerase. Gene 99(1): 105–108. https://doi.org/10.1016/0378-1119(91)90040-I

Cheung M, Au C, Chu K, Kwan H, Wong C (2010) Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. ISME Journal 4: 1053–1059. https://doi.org/10.1038/ismej.2010.26

Chou Q, Russell M, Birch DE, Raymond J, Bloch W (1992) Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Nucleic Acids Research 20(7): 1717–1723. https://doi.org/10.1093/nar/20.7.1717

Deiner K, Lopez J, Bourne S, Holman LE, Seymour M, Grey EK, Lacoursière-Roussel A, Li Y, Renshaw MA, Pfrender ME, Rius M, Bernatchez L, Lodge DM (2018) Optimising the detection of marine taxonomic richness using environmental DNA metabarcoding: the effects of filter material, pore size and extraction method. Metabarcoding and Metagenomics 2: e28963. https://doi.org/10.3897/mbmg.2.28963

Dzhembekova N, Moncheva S, Ivanova P, Slabakova N, Nagai S (2018) Biodiversity of phytoplankton cyst assemblages in surface sediments of the Black Sea based on metabarcoding. Biotechnoogy & Biotechnological Equipment 32(6): 1507–1513. https://doi.org/10.1080/13102818.2018.1532816

Dzhembekova N, Urusizaki S, Moncheva S, Ivanova P, Nagai S (2017) Applicability of massively parallel sequencing on monitoring harmful algae at Varna Bay in the Black Sea. Harmful Algae 68: 40–51. https://doi.org/10.1016/j.hal.2017.07.004

Eckert KA, Kunkel TA (1991) DNA polymerase fidelity and the polymerase chain reaction. Genome Reseach 1: 17–24. https://doi.org/10.1101/gr.1.1.17

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C, Holder M, Taylor GT, Suarez P, Varela R, Epstein S (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. The ISME Journal 5: 1344–1356. https://doi.org/10.1038/ismej.2011.6

Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B (2013) 454 Pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: A test for marine haptophytes. PLoS ONE 8(9): e74371. https://doi.org/10.1371/journal.pone.0074371

Egge ES, Johannessen TV, Andersen T, Eikrem W, Bittner L, Larsen A, Sandaa RA, Edvardsen B (2015) Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. Molecular Ecology 24(12): 3026–3042. https://doi.org/10.1111/mec.13160

Elbrecht V, Steinke D (2018) Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. Freshwater Biology 64: 380–387. https://doi.org/10.1111/fwb.13220

Fields B, Moeskjær S, Friman VP, Andersen SU, Young JPW (2021) MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. Molecular Ecology Resources 21: 703–720. https://doi.org/10.1111/1755-0998.13294

Filges S, Yamada E, Ståhlberg A, Godfrey TE (2019) Impact of polymerase fidelity on background error rates in next-generation sequencing with unique molecular identifiers/barcodes. Scientific Reports 9: e3503. https://doi.org/10.1038/s41598-019-39762-6

Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, Creer S (2012) Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Research 40(9): e66. https://doi.org/10.1093/nar/gks002

Fox J, Weisberg S, Price B, Adler D, Bates D, Baud-Bovy G, Ellison S, Firth D, Friendly M, Gorjanc G, Graves S, Heiberger R, Krivitsky P, Laboissiere R, Maechler M, Monette G, Murdoch, D, Nilsson, H, Ogle D, Ripley B, Venables W, Walker S, Winsemius D, Zeileis A, Core R (2020) Package "car": Companion to Applied Regression, 1–152. https://cran.r-project.org/web/packages/car/car.pdf [accessed 09.05.2021]

Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759- 769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Gohl DM, Vangay P, Garbe J, Maclean A, Hauge A, Becker A, Gould TJ, Clayton, JB, Johnson TJ, Hunter R Knights D, Beckman KB (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology 34(9): 942–952. https://doi.org/10.1038/nbt.3601

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlader SK, Sodergren E, Methé B, DeSantis TZ, The Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research 21: 494–504. https://doi.org/10.1101/gr.112730.110

Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR (2016) Polymerase specific error rates and profiles identified by single molecule sequencing. Mutation Research 784–785:39–45. https://doi.org/10.1016/j.mrfmmm.2016.01.003

Hirai J, Katakura S, Kasai H, Nagai S (2017a) Cryptic zooplankton diversity revealed by a metagenetic approach for monitoring metazoan communities in the coastal waters of the Okhotsk Sea, northeastern Hokkaido. Frontiers in Marine Science, 1–13. https://doi.org/10.3389/Fmars.2017.00379

Hirai J, Nagai S, Hidaka K (2017b) Evaluation of metagenetic community analysis of planktonic copepods using Illumina MiSeq: comparisons with morphological classification and metagenetic analysis using Roche 454. PLoS ONE 12(7): e0181452. https://doi.org/10.1371/journal.pone.0181452

Ishino S, Ishino Y (2014) DNA polymerases as useful reagents for biotechnology – The history of developmental research in the field. Frontiers in Microbiology 5: e456. https://doi.org/10.3389/fmicb.2014.00465

Jeunen G J, Knapp M, Spencer HG, Taylor HR, Lamare MD, Stat M, Bunce M, Gemmell NJ (2019) Species-level biodiversity assessment using marine environmental DNA metabarcoding requires protocol optimization and standardization. Ecology and Evolution 9: 1323–1335. https://doi.org/10.1002/ece3.4843

John U, Medlin LK, Groben R (2005) Development of specific rRNA probes to distinguish between geographic clades of the *Alexandrium tamarense* species complex. Journal of Plankton Research 27: 199–204. https://doi.org/10.1093/plankt/fbh160

Kalle E, Kubista M, Rensing C (2014) Multi-template polymerase chain reaction. Biomelecular Detection and Quantification 2: 11–29. https://doi.org/10.1016/j.bdq.2014.11.002

Kawato M, Yoshida T, Miya M, Tsuchida S, Nagano Y. Nomura M, Yabuki A, Fujiwara Y, Fujikura K (2021) Optimization of environmental DNA extraction and amplification methods for metabarcoding of deep-sea fish. MethodsX 8: e101238. https://doi.org/10.1016/j.mex.2021.101238

Kebschull JM, Zador AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Research 43(21): e143. https://doi.org/10.1093/nar/gkv717

Kelly RP, Shelton AO, Gallego R (2019) Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. Scientific Reports 9: e12133. https://doi.org/10.1038/s41598-019-48546-x

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq. Applied and Environmental Microbiology 79(17): 5112–5120. https://doi.org/10.1128/AEM.01043-13

Kunkel TA, Bebenek K (2000) DNA Replication Fidelity. Annual Review of Biochemistry 69: 497–529. https://doi.org/10.1146/annurev.biochem.69.1.497

Lahr DJG, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques 47(4): 857–866. https://doi.org/10.2144/000113219

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). Molecular Ecology Research 14: 753–767. https://doi.org/10.1111/1755-0998.12225

Ling LL, Keohavong P, Dias C, Thilly WG (1991) Optimization of the polymerase chain reaction with regard to fidelity: Modified T7, Taq, and Vent DNA polymerases. Genome Research 1: 63–69. https://doi.org/10.1101/gr.1.1.63

Majaneva M, Hyytiäinen K, Varvio SL, Nagai S, Blomster J (2015) Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. PLoS ONE 10(6): e0130035. https://doi.org/10.1371/journal.pone.0130035

Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenick J (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Molecular Ecology 19(Suppl. 1): 32–40. https://doi.org/10.1111/j.1365-294X.2009.04478.x

Monchy S, Grattepanche JD, Breton E, Meloci D, Sanciu G, Chabé M, Delhaes L, Viscogliosi E, Sime-Ngando T, Christaki U (2012) Microplanktonic community structure in a coastal system relative to a Phaeocystis bloom inferred from morphological and tag pyrosequencing methods. PLoS ONE 7: e39924. https://doi.org/10.1371/journal.pone.0039924

Nagai S, Chen H, Kawakami Y, Yamamoto K, Sildever S, Kanno N, Oikawa H, Yasuike M, Nakamura Y, Hongo Y, Fujiwara A, Kobayashi T, Gojobori T (2019) Monitoring of the toxic dinoflagellate *Alexandrium catenella* in Osaka Bay, Japan using a massively parallel sequencing (MPS)-based technique. Harmful Algae 89: e101660. https://doi.org/10.1016/j.hal.2019.101660

Nagai S, Hida K, Urusizaki S, Takano Y, Hongo Y, Kameda T, Abe K (2016a) Massively parallel sequencing-based survey of eukaryotic community structures in Hiroshima Bay and Ishigaki Island. Gene 576: 681–689. https://doi.org/10.1016/j.gene.2015.10.026

Nagai S, Hida K, Urushizaki S, Onitsuka G, Yasuike M, Nakamura Y, Fujiwara A, Tajimi S, Kimoto K, Kobayashi T, Ototake M (2016b) Influences of diurnal sampling bias on fixed-point monitoring of plankton biodiversity determined using a massively parallel sequencing-based technique. Gene 576: 667–675. https://doi.org/10.1016/j.gene.2015.10.025

Nagai S, Nishitani G, Tomaru Y, Sakiyama S, Kamiyama T (2008) Predation on the ciliate *Myrionecta rubra* by the toxic dinoflagellate *Dinophysis fortii* and observation of sequestration of ciliate chloroplasts. Journal of Phycology 44: 909–922. https://doi.org/10.1111/j.1529-8817.2008.00544.x

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, Shapiro B (2018) Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources 18(5): 927–939. https://doi.org/10.1111/1755-0998.12895

Nolte V, Pandey RV, Jost S (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Molecular Ecology 19: 2908–2915. https://doi.org/10.1111/j.1365-294X.2010.04669.x

Oh S, Hall R, Eng K, Hon L, Heiner C (2016). Minimization of chimera formation and substitution errrors in full-length 16S PCR amplification. PACBIO, 1 pp. https://www.pacb.com/wp-content/uploads/minimizing-chimera-formation-in-full-length-16s-pcr-amplification.pdf [accessed: 24.04.2021]

Oliver AK, Brown SP, Callaham MA, Jumpponen A (2015) Polymerase matters: Non-proofreading enzymes inflate fungal community richness estimates by up to 15%. Fungal Ecology 15: 86–89. https://doi.org/10.1016/j.funeco.2015.03.003

Ogle D, Wheeler, P, Dinno A (2021) Package "FSA". 203 pp. https://cran.r-project.org/web/packages/FSA/FSA.pdf [accessed 09.05.2021]

Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, Zhou J (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Applied Environmental Microbiology 67: 880–887. https://doi.org/10.1128/AEM.67.2.880-887.2001

Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, Mcquillan JA, Swerdlow HP, Oyola SO (2012) Optimal enzymes for amplifying sequencing libraries. Nature Methods 9(1): 10–11. https://doi.org/10.1038/nmeth.1814

Patin NV, Kunin V, Lidström U, Ashby MN (2013) Effects of OTU clustering and PCR artifacts on microbial diversity estimates. Microbial Ecology 65: 709–719. https://doi.org/10.1007/s00248-012-0145-4

Pfeifer SP (2017) From next-generation resequencing reads to a high-quality variant data set. Heredity 118: 111–124. https://doi.org/10.1038/hdy.2016.102

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports 8: e10950. https://doi.org/10.1038/s41598-018-29325-6

Potapov V, Ong JL (2017) Examining sources of error in PCR by single-molecule sequencing. PLoS ONE 12(1): e0169774. https://doi.org/10.1371/journal.pone.0169774

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: e38. https://doi.org/10.1186/1471-2105-12-38

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ [accessed 10.04.2021]

Rhodes LL (1998) Identification of potentially toxic *Pseudo-nitzschia* (Bacillariophyceae) in New Zealand coastal waters, using lectins. New Zealand Journal of Marine and Freshwater Research 32: 537–544. https://doi.org/10.1080/00288330.1998.9516842

Santoferrara L (2019) Current practice in plankton metabarcoding: Optimization and error management. Journal Plankton Research 41(5): 572–582. https://doi.org/10.1093/plankt/fbz041

Sawaya NA, Djurhuus A, Closek CJ, Hepner M, Olesin E, Visser L, Kelble C, Hubbard K (2019) Assessing eukaryotic biodiversity in the Florida Keys National Marine Sanctuary through environmental DNA metabarcoding. Ecology and Evolution 9: 1029–1040. https://doi.org/10.1002/ece3.4742

Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 6: e27310. https://doi.org/10.1371/journal.pone.0027310

Sildever S, Kawakami Y, Kanno N, Kasai H, Shiomoto A, Katakura S, Nagai S (2019) Toxic HAB species from the Sea of Okhotsk detected by a metagenetic approach, seasonality and environmental drivers. Harmful Algae 87: e101631 https://doi.org/10.1016/j.hal.2019.101631

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular Ecology 19(Suppl. 1): 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Sze MA, Schloss PD (2019) The Impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. mSphere 4: e00163–19. https://doi.org/10.1128/mSphere.00163-19

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA for biodiversity research and monitoring. Oxford University Press, 43–44. https://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Takagi M, Nishioka M, Kakihara H, Kitabayashi M, Inoue H, Kawakami B, Oka M, Imanaka T (1997) Characterization of DNA polymerase from *Pyrococcus* sp. strain KOD1 and its application to PCR. Applied Environmental Microbiology 63(11): 4504–4510. https://doi.org/10.1128/aem.63.11.4504-4510.1997

Tanabe AS, Satoshi N, Kohsuke H, Motoshige Y, Atushi F, Yoji N, Yoshihito T, Seiji K (2015) Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. Molecular Ecology Resources 16: 402–414. https://doi.org/10.1111/1755-0998.12459

ThermoFisher Scientific (2021) Performance Comparison of Platinum SuperFi II and Platinum SuperFi DNA Polymerases, 1–14. https://assets.thermofisher.com/TFS-Assets/BID/Reference-Materials/platinum-superfi-ii-and-superfi-benchmarking-data.pdf [accessed: 13.05.2021]

van der Loos LM, Nijland R (2020) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology 00: 1–19. https://doi.org/10.1111/mec.15592

van Pelt-Verkuil E, van Belkum A, Hays JP (2008) Principles and technical aspects of PCR amplification. Springer-Verlag, Berlin, 325 pp.

Wang Y, Prosen DE, Mei L, Sullivan JC, Finney M, Vander Horn PB (2004) A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance *in vitro*. Nucleic Acids Research 32(3): 1197–1207. https://doi.org/10.1093/nar/gkh271

Wintzingerode FV, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews 21: 213–229. https://doi.org/10.1111/j.1574-6976.1997.tb00351.x

Wright ES, Yilmaz LS, Noguera DR (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Applied Environmental Microbiology 78(3): 717–725. https://doi.org/10.1128/AEM.06516-11

Zaiko A, Greenfield P, Abbott C, von Ammon U, Bilewitch J, Bunce M, Cristescu ME, Chariton A, Dowle E, Geller J, Ardura Gutierrez A, Hajibabaei M, Haggard E, Inglis GJ, Lavery SD, Samuiloviene A, Simpson T, Stat M, Stephenson S, Sutherland J, Thakur V, Westfall K, Wood SA, Wright M, Zhang G, Pochon X (2021) Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment. Molecular Ecology Resources 00: 1–20. https://doi.org/10.1111/1755-0998.13485

Zhu XJ, Sun S, Xie B, Hu X, Zhang Z, Qiu M, Dai ZM (2016) Guanine-rich sequences inhibit proofreading DNA polymerases. Scientific Reports 6: e28769. https://doi.org/10.1038/srep28769

Supplementary material 1
Table S1–S3

Author: Nagai S, Sildever S, Nishi N, Tazawa S, Leila B, Kobayashi T, Ishino Y

Data type: accession numbers

Explanation note: **Table S1.** Information source and accession numbers on the nuclear ribosomal RNA gene (18S rRNA) of 40 species used for the mock sample. ND: no data. **Table S2.** Composition of PCR mixtures and PCR conditions. PCR amplification was done by using 14 PCR kits, for selected kits 3 different annealing temperatures were tested. **Table S3.** P-values resulting from posthoc comparisons of differences among PCR kits containing polymerase based on the seven parameters (A-G). Statistically significant p-values after Benjamini-Hochberg correction are marked with bold and in italics.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl1

Supplementary material 2
merge_tophit_count.pl

Author: Nagai S, Sildever S, Nishi N, Tazawa S, Leila B, Kobayashi T, Ishino Y

Data type: Blast XML

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl2

Supplementary material 3
same_tophit_count_merge.pl

Author: Nagai S, Sildever S, Nishi N, Tazawa S, Leila B, Kobayashi T, Ishino Y

Data type: Blast XML

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl3

Supplementary material 4
blastxml_parser

Author: Nagai S, Sildever S, Nishi N, Tazawa S, Leila B, Kobayashi T, Ishino Y

Data type: Blast XML

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl4